REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ADDRESS.

1. REPORT DATE *(DD-MM-YYYY)*	2. REPORT TYPE	3. DATES COVERED *(From - To)*
12-21-2012	Final	November 2010 - November 2012

4. TITLE AND SUBTITLE

Proactive Response to Potential Material Shortages Arising from Environmental Restrictions Using Automatic Discovery and Extraction of Information from Technical Documents

5a. CONTRACT NUMBER

SP4701-10-C-0018

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)

Dr. David F. Winchell, Ph.D.

5d. PROJECT NUMBER

4516057545

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

XSB, Inc
21 Bennetts Road Suite 100
Setauket, New York 11733

8. PERFORMING ORGANIZATION REPORT NUMBER

IBIFA002

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

DLA Contracting Services Office Philadelphia
700 Robbins Avenue
Attn:DSCO-P 26-1 J. Dormer
Philadelphia, PA 19111
Email: John.Dormer@dla.mil

10. SPONSOR/MONITOR'S ACRONYM(S)

HQ-DLA J-332

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT

Approved for public release; distribution is unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT

We report on the methods and results of the Strategic Materials Assessment and Research Tool (SMART). This project was undertaken to help provide a proactive response to potential material shortages in a dynamic regulatory environment. Existing technologies were extended, and new ones developed, with the goal of discovering, extracting, and disseminating information regarding materials and regulatory compliance for items in the supply chain. The tools were used to extract the relevant information from a wide variety of sources, and an interface was developed for the dissemination of the information. Prime and secondary contractors were involved in evaluating the results of the project for two weapons systems.

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			Daria Wesnofske
0	0	0	None	24	19b. TELEPHONE NUMBER *(Include area code)* 631-371-8102

Reset

Proactive Response to Potential Material Shortages Arising from Environmental Restrictions Using Automatic Discovery and Extraction of Information from Technical Documents

Contract Number:
SP4701-10-C-0018

Performing Organization:
XSB, Inc.
Dr. David Winchell
21 Bennetts Road- Suite 100
Setauket, NY 11733
Voice: 631-371-8117
Fax: 631-382-8228
Email: d.winchell@xsb.com
Effective Date of Contract: 17 November 2010
Reporting Period: November 2010- November 2012

Issuing Organization:
DLA Contracting Services Office Philadelphia
700 Robbins Avenue
Attn:DSCO-P 26-1 J. Dormer
Philadelphia, PA 19111
Email: John.Dormer@dla.mil

Proactive Response to Potential Material Shortages Arising from Environmental Restrictions Using Automatic Discovery and Extraction of Information from Technical Documents
Final Scientific and Technical Report
Contract No. SP4701-10-C-0018
XSB, Inc. – Dr. David Winchell

Abstract

We report on the methods and results of the Strategic Materials Assessment and Research Tool (SMART). This project was undertaken to help provide a proactive response to potential material shortages in a dynamic regulatory environment. Existing technologies were extended, and new ones developed, with the goal of discovering, extracting, and disseminating information regarding materials and regulatory compliance for items in the supply chain. The tools were used to extract the relevant information from a wide variety of sources, and an interface was developed for the dissemination of the information. Prime and secondary contractors were involved in evaluating the results of the project for two weapons systems.

Proactive Response to Potential Material Shortages Arising from Environmental Restrictions Using Automatic Discovery and Extraction of Information from Technical Documents
Final Scientific and Technical Report
Contract No. SP4701-10-C-0018
XSB, Inc. – Dr. David Winchell

Table of Contents

Lists of Figures and Tables

Figures

Tables

Proactive Response to Potential Material Shortages Arising from Environmental Restrictions Using Automatic Discovery and Extraction of Information from Technical Documents
Final Scientific and Technical Report
Contract No. SP4701-10-C-0018
XSB, Inc. – Dr. David Winchell

Summary

Changes to international regulations regarding materials can have far-reaching impacts on the defense supply chain that are not immediately apparent, and responses to these changes tend to be reactive rather than proactive. In order to develop a proactive stance towards an evolving regulatory environment, it is necessary to have as much accurate, timely, and relevant information as possible about items in the supply chain. Often, this information is dispersed in a number of documents of different types. In particular it is desirable to have as complete a picture of constituent materials as possible. Additional information can be inferred from manufacturer compliance statements regarding parts or classes of parts.

In order to collect material and compliance information from a variety of sources, new technologies for information extraction were combined with existing tools. Two key technologies developed or extended for this project were the extraction of tagged text from PDF documents, and custom annotators for finding material data and other key information in a UIMA environment.

In the course of this project, the tools and methods developed were used to extract and aggregate material and regulatory information for over one million items. Of these, over 11,000 where found to contain materials on the REACH Substances of Very High Concern (SVHC) list.

In order to evaluate and validate the results of the extraction, we collaborated with Northrop Grumman Corporation (NGC). The extraction results for parts from two weapons systems where compared with in-house information regarding material constituents.

The results of the information extraction from this project were made available through the Pin Point web portal, which is freely available to Defense Department personnel. Development of the web interface was carried out in collaboration with personnel from the Defense Logistics Agency.

Proactive Response to Potential Material Shortages Arising from Environmental Restrictions Using Automatic Discovery and Extraction of Information from Technical Documents
Final Scientific and Technical Report
Contract No. SP4701-10-C-0018
XSB, Inc. – Dr. David Winchell

Introduction

In an increasingly connected and regulated world economy, material regulations in one country or group of countries can significantly affect product availability in other countries. In particular, changes in technologies made by American companies in response to material regulations in the European Union or China can impact the availability of parts to support DOD weapon systems. A well known example of this is the use of lead free solder in electronic components in response to Restrictions of Hazardous Substances (RoHS) promulgated by the EU. This substitution was made without regard to the unsuitability of these lead free solder formulations in military space applications. In general, this situation can lead to significant diminishing manufacturing sources and material shortage (DMSMS) problems in the military supply chain.

To date, DMSMS has been in a reactive mode to these regulatory changes. Responding to changes in material or process technology necessitated by regulatory compliance occurs after these technology changes are implemented and problems with military part applications are revealed. This situation is becoming more significant as an increasing number of material restrictions are added to regulatory programs such as the European Union Registration, Evaluation and Authorization of Chemicals (REACH) program. REACH was first instituted in January of 2007 and every six months additional materials are added to the REACH regulatory environment.

The technical goal of this project is to identify, retrieve, and extract relevant information from electronic documents pertaining to the material composition of weapon system components. The documents can take a variety of forms, including product datasheets, Material Safety Data Sheets, or manufacturer-specific environmental compliance documents.

Methods, Assumptions, and Procedures

Overview

One of the key aspects of this project is the extraction and collection of data from a variety of sources. The extraction techniques developed for this project were implemented as processing chains to move the data from original document to final database. The figure below illustrates a top-level view of the combined processing, with different paths for data extracted from the various sources.

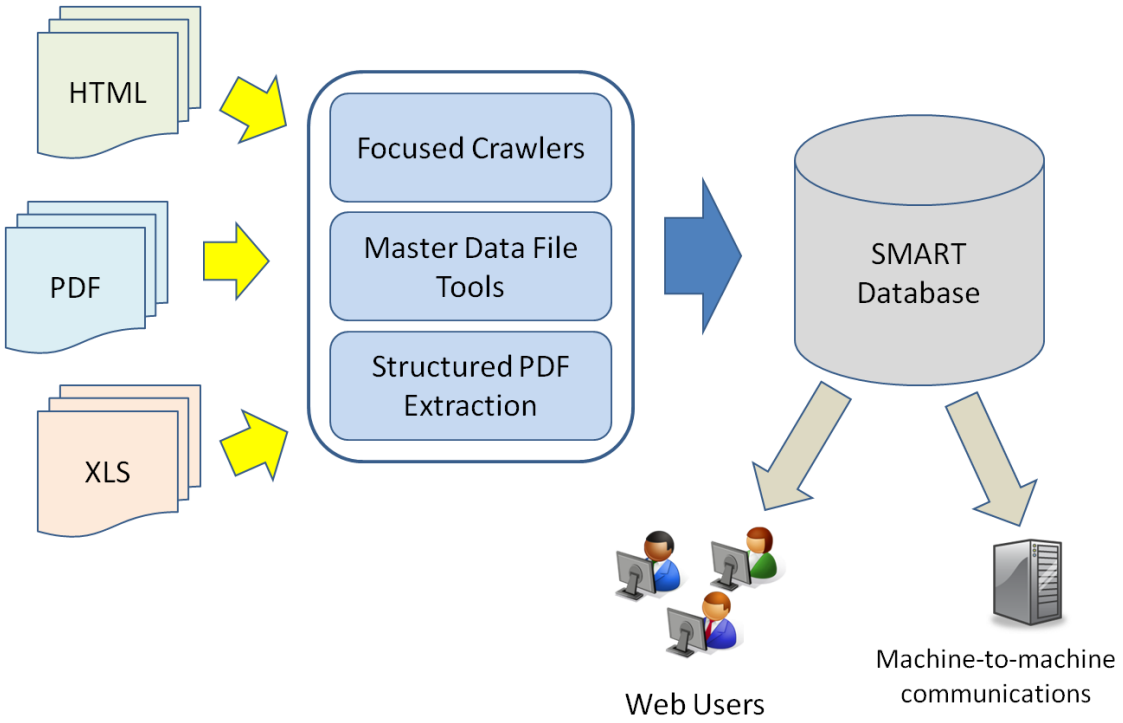


Figure 1 - Overview of data collection processes

In practice, there is more than one processing path for a given input type. For instance, for PDF documents slightly different procedures were used for Material Safety Data Sheets (MSDS) and Product Data Sheets (PDS). In a few cases, custom extraction paths were developed for small, data-rich sources.

Database Design and Population

A key element of this effort is the database used to aggregate data extracted from a variety of sources. Among these data sources are:

- Bill-of-material data regarding target weapon systems from Northrop Grumman and subcontractors
- CAS number / Material name relationships from a variety of public sources

- Lists of materials of concern such as REACH Substances of Very High Concern (SVHC), DoD emerging contaminants, and others
- Coherent View data for matching parts with National Stock Numbers (NSNs)
- Material information extracted from MSDS documents
- Material information extracted from product datasheets
- Compliance statements from manufacturers and distributors
- Material and compliance data from crawled web sites

For incoming data, the table structures reflect the nature of the data being collected. For instance, Northrop Grumman provided bill of material data for the Apache Longbow Fire Control Radar (LBFCR) and Joint Strike Fighter Distributed Aperture System (JSFDAS) in Excel format. The information provided included part numbers, CAGE codes, and short descriptions. In addition, there were in some cases tables for cross-referencing NGC internal part numbers with manufacturer names and part numbers.

The information for associating CAS registry numbers with materials came largely from the US government Substance Registry Services (SRS)[1]. For this data, a comma-separated list of material names, synonyms, and CAS numbers was obtained through the Data.gov web site.

For material data extracted from PDF documents, web pages, and other sources, it was critical to keep track of the source information, the extraction process, and the detailed material data. In general, for this type of data, two tables where created for each type of source: one containing source and extraction process information, and one containing detailed material information. For instance, for MSDS files, one table contains file name, product name and manufacturer. For each file, a unique numeric identifier is defined. In the second table, that identifier is associated with the material information extracted from that file. The material information includes material name and CAS number, and percent weight where available.

For crawled data, the source information is taken from the master table for crawled data. This table includes product name and manufacturer, the crawl job identifier (where a crawl job is defined by a starting URL and a date), along with a unique numeric identifier for the item. In the crawl database, each item has several data records associated to it, which contain attribute/value pairs and source URLs. For the purposes of this project, we filtered this data to pull in only material-related and regulation-related data records. One aspect of the data records from the crawl database is that they are not normalized; attributes and values are taken directly from the web page. In light of this, additional tables were needed to contain the normalized material and regulatory information from the crawl data. The normalization procedures will be discussed elsewhere in this report.

[1] http://ofmpub.epa.gov/sor_internet/registry/substreg/home/overview/home.do

Ultimately, we need to aggregate the information from all the sources into common data structures. Specifically, we want to have one table containing all material information, and another for all regulatory information. In each case, the item is identified by numeric identifier and "catalog", where catalog refers to MSDS documents, crawled items, etc. In addition, we have a source table, where each data source (document name, web URL, etc) is assigned an identifier.

Finally, we need to have data structures in place to relate identical items between or within catalogs. For this we followed a standard in-house process known as "master data file" (MDF) generation. There are two key MDF tables: one to associate a "group number" to a specific, standardized manufacturer name and part number, and another to associate these group numbers with the catalog name/item number pairs in the extracted data.

Document Discovery and Acquisition

From the beginning of the project, it was expected that a good deal of the detailed material information would come from MSDS documents. One source of these documents is via web links on manufacturer and distributor product catalogs on the web. These links were discovered using XSB's focused crawler technology, which will be discussed further below. We also obtained a collection of MSDS documents through a commercial source, Hazcom MSDS.

Product data sheets in PDF format can also contain information on constituent materials for products, especially for electronics components, which do not typically have associated MSDS documents. As with MSDS, links to product datasheets were discovered via the focused crawler.

In addition to PDF documents, manufacturers and distributors often include material information in the descriptions of items in on-line catalogs. Using our existing focused crawler technology, we collect part number/attribute/value "triples" from these catalogs. The extraction of detailed material information from this data will be described below. In cases where the web page points to a PDF document, we collect the link URL as the value in the triple, and the link text as the attribute. Using this, we can collect links to these documents, as mentioned above, and download the files in an automated way.

In some cases, information of interest was contained in "one-of-a-kind" files that didn't lend themselves to mass acquisition, but were worthwhile harvesting by hand. These include corporate statements of compliance, and Excel spreadsheets containing information on REACH SVHC and other environmentally sensitive materials.

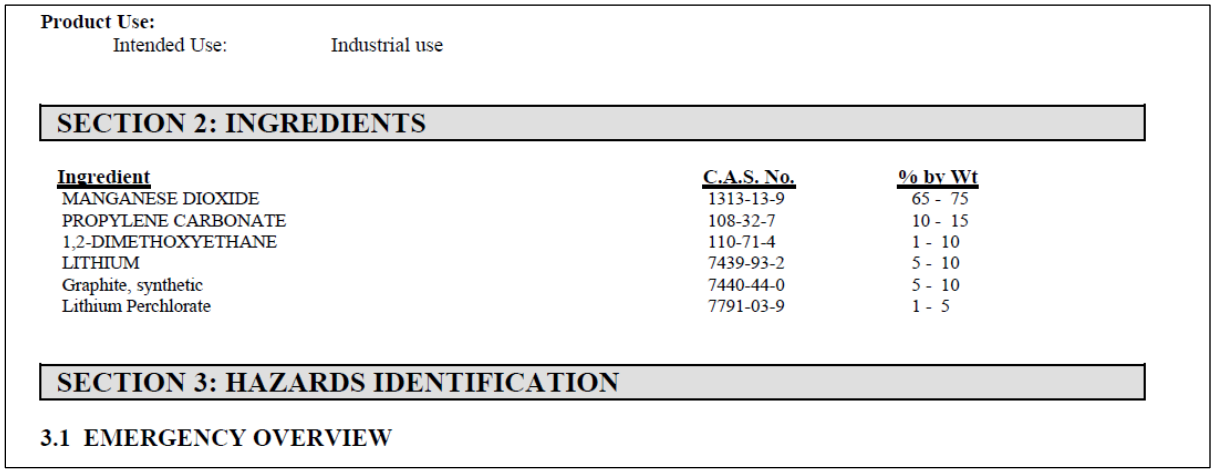
Data extraction from PDF Documents

Conversion to tagged text

One of the primary goals of this project was the development of software to automatically discover and extract information from PDF files. The task is challenging because these files are designed to provide human-readable documents, rather than machine-to-machine communication. Within the file, each word of text, or in many cases each single character, is associated with a position on the page. In order to extract meaningful information, we need to put these text elements together into a logical structure. It is not sufficient to simply "scrape" the text from the file; in order to extract meaningful information we need to preserve or reconstruct information about structures such as tables, lists, section titles, *etc*.

To build software to carry this out, we began with the open-source PDFBox library from the Apache Software Foundation. This library includes software that will read a PDF file and provide information on the position, orientation, font, and size of every character on the page. We then take that information and reconstruct words based on character spacing. Using this information, we analyze each page of the document, and identify header and footer information. Next, white space and graphical elements such as lines on the remaining portion of each page are analyzed to identify tables.

Figure 2 shows a section from an MSDS document containing a table of ingredients. In order to correctly identify and tag the table and table elements such as rows and columns, our software analyzes each block of text on the page, and identifies contiguous sections of whitespace, as exist between the columns in the figure. The software also identifies contiguous blocks of left-justified or center-aligned text, as often occurs in table columns. Using this and other features, all likely combinations of text blocks are examined and scored and the combinations that look most like tables are tagged.



Product Use:
Intended Use: Industrial use

SECTION 2: INGREDIENTS

Ingredient	C.A.S. No.	% by Wt
MANGANESE DIOXIDE	1313-13-9	65 - 75
PROPYLENE CARBONATE	108-32-7	10 - 15
1,2-DIMETHOXYETHANE	110-71-4	1 - 10
LITHIUM	7439-93-2	5 - 10
Graphite, synthetic	7440-44-0	5 - 10
Lithium Perchlorate	7791-03-9	1 - 5

SECTION 3: HAZARDS IDENTIFICATION

3.1 EMERGENCY OVERVIEW

Figure 2 - Section MSDS document showing materials table

The end result of the process is a plain-text version of the document with tags to indicate structural features such as tables, lists, and page headers.

Data extraction using the Unstructured Information Management Architecture

Once the tagged text files are created, we run custom code to extract information of interest. This information includes materials, part numbers and company names, and regulatory compliance information. In order to carry this out, we use the Unstructured Information Management Architecture (UIMA) library from the Apache Software Foundation. Using this architecture, a given document is run through several "annotators" to find specific types of information within the text. Each annotator can act on the document contents as well as the results from previous annotators. At the end of the process, the cumulative results from the annotators are used to extract information from the document.

For this project, several custom annotators were written and used. Most importantly, there were several annotators that helped to recognize information on materials. Towards this end, there were three key material annotators in use.

First, we adapted existing technology written in the XSB Prolog programming language, using a material taxonomy developed for the XSB, Inc. Coherent View™ database. In order to integrate this technology into the Java-based UIMA annotation system, we built a web service to allow the Prolog process to run on a separate machine and be invoked when needed. The Java annotator passes a section of text to the web service, which applies the Prolog-based extraction and passes back information on materials found in the text. The existing Prolog extraction tool was extended to provide information about the position of the extracted information in the original text, as this positional information is needed by the UIMA framework. In addition, several dozen new materials were added to the taxonomy used in this process, in order to provide coverage of the materials found in the Reach SVHC list.

The second key material annotator developed for this project involved recognizing words and word groups that are likely chemical names. This is a two-step process. First, words within the document are matched to patterns that are likely to be found in chemical names, such as the prefix "dioxy". Once these words are tagged, neighboring words are analyzed to determine if a group of adjacent words constitute a material name.

The third method makes use of a standard table pattern in MSDS documents. Using the tagged text file, it finds the document section and table containing information on hazardous materials. Typically, this table will include columns for material name, CAS number, and percent weight.

Figure 3 shows an example of tagged text. In this example, chemical names and CAS numbers found in a table in an MSDS document are highlighted. Once the annotation have been created, the data is collected into a two database tables, one associating part information with the document, and the other including material names, CAS numbers, and percentage amounts. Some of the other annotators

can be seen in the color-coded list in the bottom portion of the figure. These include product name, document creation date, and manufacturer.

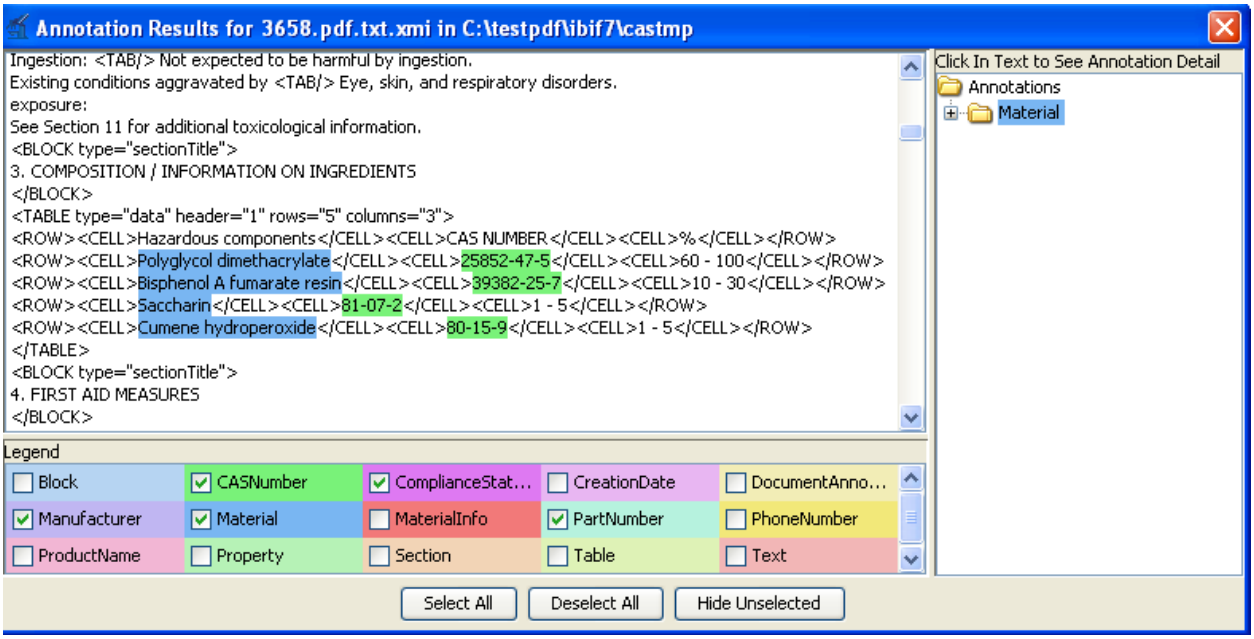


Figure 3 - Example of annotated text

Along with material information, annotators were written to identify words or phrases indicating regulatory compliance, such as “RoHS Compliant”, which often occur in product data sheets.

Finally, annotators were written to help identify part numbers and company names, in order to associate material and regulatory information with specific items

Focused Crawler Data

In addition to PDF documents, we made use of data collected from the web using XSB, Inc.’s focused crawler technology. This technology collects pages from online product catalogs and automatically extracts attribute information for items listed in the catalog. As a rule, the focused crawler extracts attributes and values as they appear on the web page. To be useful for this project, it was necessary to normalize the data by extracting standardized material names from the value fields. For this, we used the same Prolog-based material extraction mentioned in the previous section. First, attributes were filtered based on patterns that would indicate material composition, such as “material”, “ingredient”, “contains”, “plating”, etc. The values associated with these attributes are then run through the material extraction web service discussed above, to normalize material names and to extract multiple materials from value fields containing more than one.

Aside from data on HTML web pages, some of the PDF documents analyzed in this project were obtained through the focused crawler. In these cases, the detailed part number and manufacturer information obtained in the crawl process could be used instead of the item information obtained through UIMA annotation.

Custom Extractions

The processes used for information extraction from PDF documents and web sites as described above were designed to be scalable, so that large numbers of documents can be processed with minimal user intervention. There were also a few data sources used in this project for which custom extraction procedures were written. The most important of these were extraction from an Excel spreadsheet for compliance data from Vishay, and extraction of material information from XML files from the web site of Fairchild Semiconductor.

Part Number / Part Series Resolution

For some data sources, information was associated with parts series rather than individual parts. In order to make the association to individual parts, we developed database structures to associate series names and numbers with specific part numbers. In general, the rules for relating parts to series information was done on a case-by-case basis and is not as scalable as most other procedures developed in this project.

Master Data File Technology

In order to normalize parts information from a variety of sources, we used the existing XSB automated Master Data File (MDF) generation process. This process creates clean, structured parts data by finding identical items based on normalized part numbers and manufacturer names. This system makes use of an extensive knowledge base of company name variations, and of relations between company names arising from mergers and acquisitions. For example, the chromate conversion coating Alodine 1200 was made by Amchem Products, Inc., which has become part of Henkel. Our system recognizes it as the same item regardless of which manufacturer is listed.

Comparing Extraction Results to Weapon Systems Data

As mentioned above, a portion of this effort was carried out in collaboration with the Northrop Grumman Corporation (NGC), in order to evaluate our data collection results using corporate information on two weapons systems. To carry this out, NGC supplied us with part numbers and descriptive data for the constituent parts in these systems. We then shared our results with them for further evaluation. NGC also solicited information from two of their suppliers, DuPont and Henkel, for inclusion in the evaluation. The results of these evaluations will be discussed in the following section.

Update Procedures

In order for the data to be useful in the long term, it is necessary that regular updates be made to both parts information and to regulatory data. This was carried out by creating automated procedures for the various extraction processes. For example, one procedure will take as input a directory containing a number of MSDS files in PDF format. The procedure will run the programs necessary to generate tagged text, extract data, and write that data to a database.

Pin Point Integration

In order to provide access to the data collected in this project, the extracted material and compliance information were integrated into the Pin Point web application (pinpoint.xsb.com) which provides access to Coherent View™ (CV) data and is available to all Defense Department employees.

The CV data is updated on a quarterly basis. This update involves the collection and processing of data from a number of sources. As part of this project, the quarterly update procedures were modified to pull in the data from this project on materials and regulations, and merge it into the CV databases.

To disseminate the data, forms were added to the Pin Point web application for retrieving information based on lists of parts input by the user. After the initial development of the interface, we carried out training and had a two-month beta-test period involving personnel from the Defense Logistics Agency. Based on feedback from this period, several improvements to the interface were carried out. Further discussion of the interface and its capabilities is provided in the next section.

Results & Discussion

Overall Results for Material Information

As discussed above, the two primary sources of parts data for this project were MSDS documents and web pages obtained using the Focused Crawler. In addition, we analyzed data supplied by Northrop Grumman for the two weapon systems of interest. Finally, some data was collected via custom extractions from Fairchild (XML documents obtained by crawler), Vishay (product data sheets), Omron (product data sheets), and Texas instruments (HTML documents from crawler with complex tables requiring special extraction).

It should be noted that, in the dissemination phase of this project, the Northrop Grumman data was not exported to the public interface. Rather, it was used as a metric to help determine the success of the extraction processes. This will be discussed further below.

The following table shows, by data source, the number of items found, the total number of distinct items after applying the MDF process, and the number of distinct items for which material information was extracted. For each source, the number of distinct items is smaller than the total number of items due to multiple instances, for instance two web catalogs selling the same item. As can be seen from the table, the percentage of items for which material data is found can vary from source to source. For the focused crawler data, material information was found for about 29% of the distinct items, and for MSDS data material information was found for about 84% of the items.

Table 1 - Items by Source

Source	Total items	Distinct Items	Distinct Items w/material
Focused Crawler	3986696	3357096	965853
MSDS	305845	236057	198084
NGC	5818	5505	2585
Fairchild XML	8036	8016	8016
TI Crawled	2172	2172	1023
Vishay PDS	304	304	304
Omron PDS	1208	1207	1207

Overall Results for Compliance Information

As with material data, a fair amount of regulatory compliance information was found in extractions from focused crawler data and custom data feeds. While MSDS documents don't generally have this kind of information, we were able to extract some from PDF product data sheets and manufacturer-supplied spreadsheets. Finally, we inferred compliance for some items based on blanket statements provided by

manufacturers (i.e. “all items in this catalog are REACH compliant …”). Table 2 shows the number of unique items with compliance information by data source, along with the types of regulations covered (Note: ELV = End-of-Life Vehicle Directive). Because an item may have compliance statements for more than one regulation, the number in the second column may be less than the sum of the last four columns.

Clearly, different types of information are collected from different documents. Statements about REACH compliance are more likely to be given in blanket compliance statements or manufacturer spreadsheets, while RoHS information is likely to be associated with individual items on web pages or product datasheets.

Table 2 - Type and source of compliance information

Source	Items	REACH	RoHS	Lead-Free	ELV
Crawled Data	287492	-	280716	75025	143349
Compliance Statements	63649	63649	204	-	-
Fairchild XML	5053	-	5053	-	-
Product Datasheets	5284	-	5275	1501	188
Manufacturer spreadsheets	5773	3942	1831	-	-

Results for Specific Northrop Grumman Weapons Systems

One of the goals of this project is to compare our results with specific weapons systems built by Northrop Grumman Corporation (NGC), namely the Apache Longbow Fire Control Radar (LBFCR) and the Joint Strike Fighter Distributed Aperture System (JSFDAS). NGC supplied bills of material for these systems, which included part number and manufacturer information, and in some cases included brief textual descriptions.

Combining the bills of material, there were 5505 distinct parts. Of these, 181 were found in our focused crawler data, 27 were found in the data collected from MSDS documents, and 25 were found in the specialized extraction done on Fairchild XML documents. From this data, we found 995 material records. Of these, 17 records matched materials on the REACH SVHC list. The SVHC materials found were dibutyl phthalate, N-methyl-2-pyrrolidone, chromic acid, and chromium trioxide.

For completeness, we also extracted material information from the text records supplied with the NGC bills of material. From these, we extracted 3225 material records for 2104 items. Included in this were three occurrences of strontium chromate, which was added to the SVHC list in 2011. It should be noted, however, that information extracted from NGC-supplied data is considered proprietary and was not included in the dissemination to Pin Point described below.

In addition to material information, 333 compliance statements for 144 items were extracted from focused crawler data, compliance statements, or manufacturer spreadsheets matching the NGC parts. Of these, 141 statements indicated REACH compliance for the items.

In order to validate our results, NGC evaluated our extracted data for the LBFCR and JSFDAS against their internal documents and information obtained from key suppliers. Their results are detailed in the attached Subcontractor Final Report.

Tagged Text from PDF Documents

On the whole, the extraction of tagged text from PDF documents was successful. We were able to recognize and tag document tables. This was especially true for MSDS documents, where table structures tended to be more regular than in product data sheets. The automatic recognition of table headers and footers, and of section headings, was also adequate to our purposes.

Information Extraction from Tagged Text

Applying the UIMA annotation and extraction process was successful for materials and CAS numbers. The primary problems that arose had to do with filtering the extraction results by context. For example, a product data sheet for adhesives or tapes might indicate that the product “adheres to steel”. In this case the annotator would recognize “steel”, even though we don’t want to associate that material as a constituent of the item. For MSDS documents, because we could almost always successfully tag section information, we could restrict the extraction to “component materials” or “hazardous materials” sections. For product data sheets, we tried several strategies to recognize context, but the overall results contained too many false extractions to be useful, and the results were not added to the database. We did eventually collect material information for about 1500 items from product datasheets. These were cases where we were able to build custom extractors in order to exploit particular table structures. Unfortunately, this solution is not particularly scalable.

The other issue that arose in the UIMA-based information extraction was the identification of product name/part number/manufacturer data. Many documents have clear labels or other indicators to show which text corresponds to item identification information. However, in many other cases there were not clear indicators and identification information was not retrieved or was extracted incorrectly. For documents collected in conjunction with focused crawlers (a link to the PDF document was collected as an attribute by the crawler) we could associate product identification information with the document based on results from the crawler. This was done for slightly over eighteen thousand items in the course of this project.

Dissemination through Pin Point

In order to disseminate the information collected in this project, dedicated pages within the Pin Point web portal were created. As mentioned above, many of the features included in this interface were suggested by a beta-test group working with an early version of the interface. There are two forms for retrieving the data, one based on part numbers or NSNs, the other based on material queries. In addition, we created a subscription service that allows users to register a list of part numbers or NSNs, and receive alerts if a material associated with those parts is added to the REACH SVHC list.

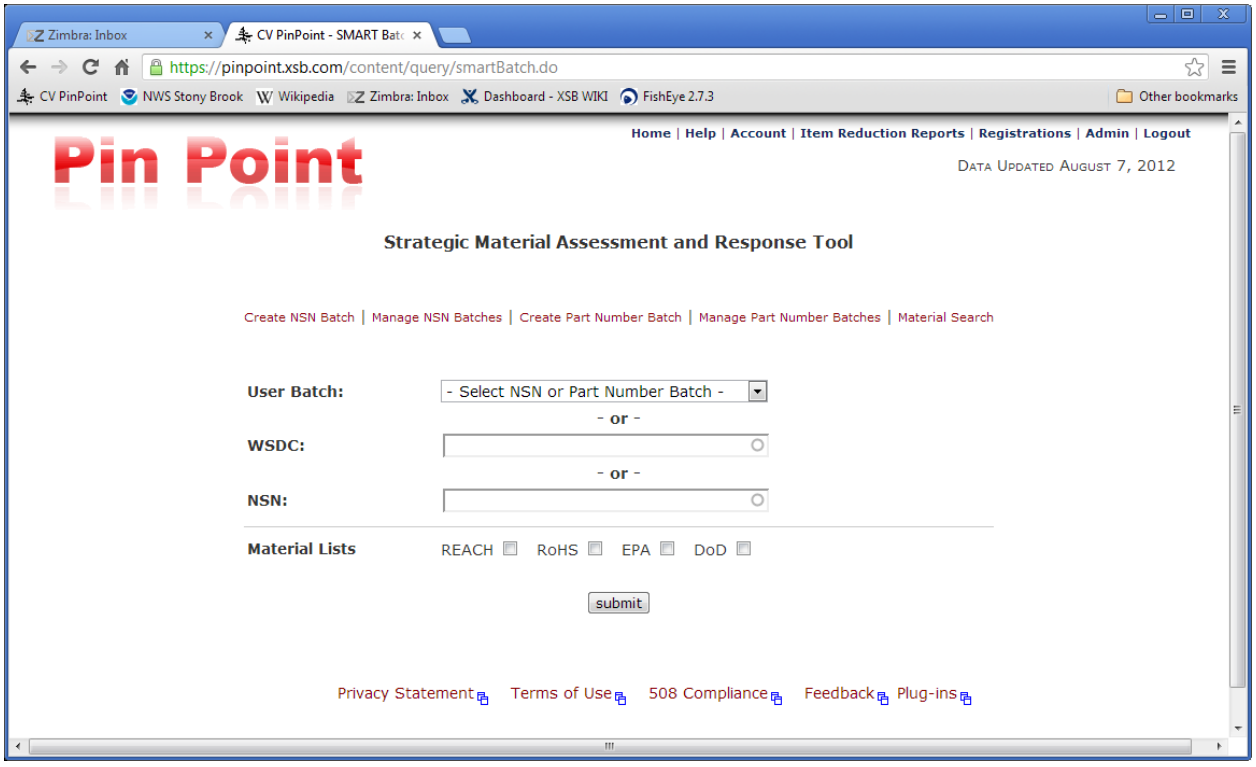


Figure 4 - Main SMART search form in Pin Point

Figure 4 shows a snapshot the parts-based query screen. The user can define a list of parts, or a list of NSNs, or a single NSN. The user can also specify a weapon systems designator code (WSDC), which will search over a list of all NSNs associated with that system. The user can also choose one or more of the four “lists of concern” (RoHS, REACH SVHC, EPA Priority, or DoD Emerging Contaminants).

When the user makes his or her choices and clicks “submit”, two tables of information are returned. First, as table associating materials to parts will be presented. This table includes fields for part number, company name, NSN, Status Code, material, CAS Number, and percent weight. It also contains columns

indicating which lists of concern the material appears on. The second table shows all extracted compliance information for the requested parts. Figure 5 and Figure 6 show examples of retrievals for WSDC = 05F (AIRCRAFT, STRATOLIFTER C/KC-135).

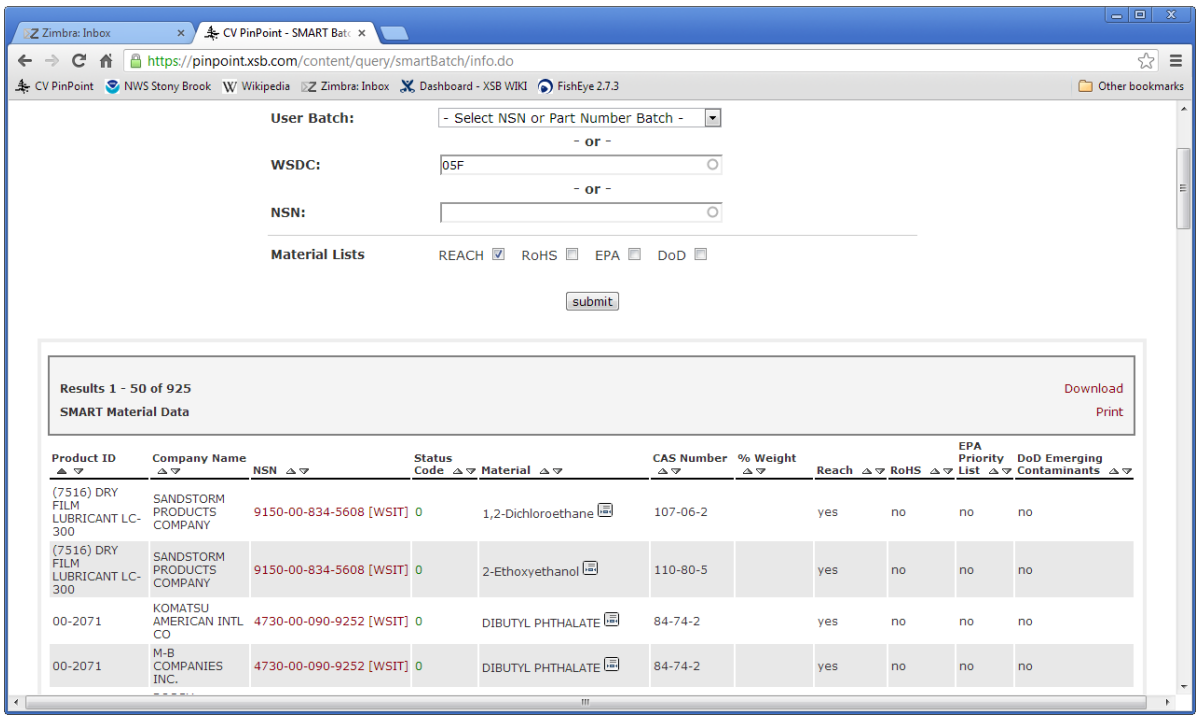
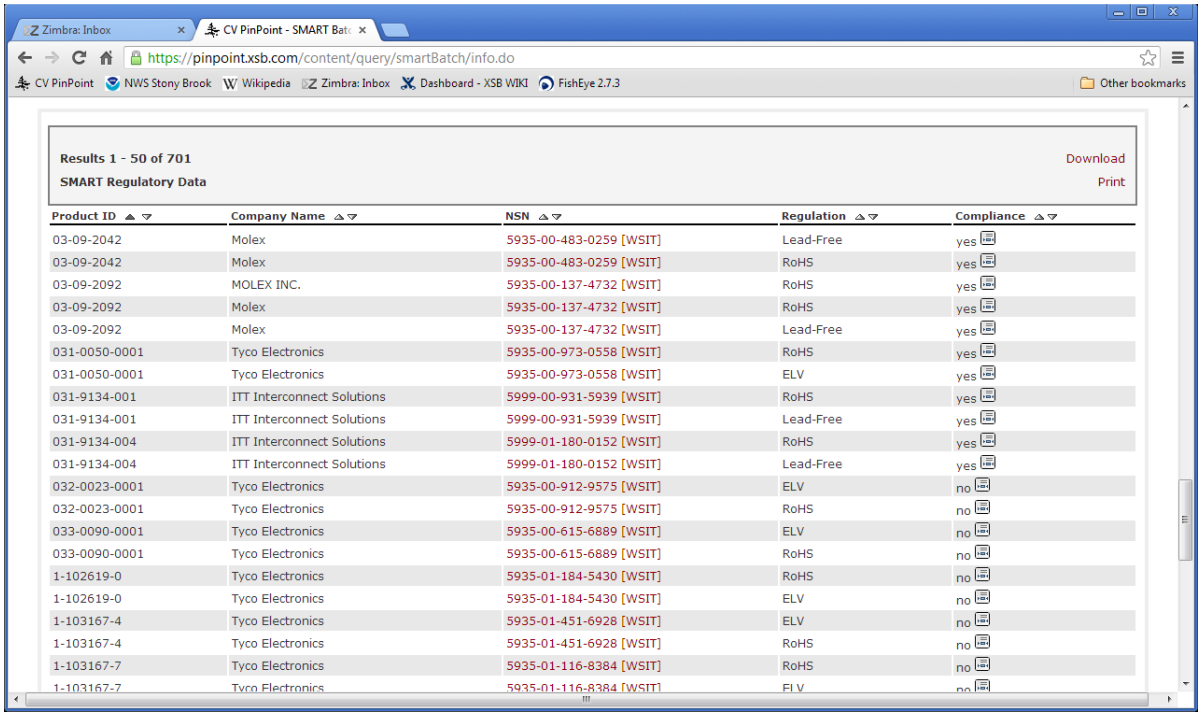


Figure 5 - Materials search results for specified weapons system



Product ID	Company Name	NSN	Regulation	Compliance
03-09-2042	Molex	5935-00-483-0259 [WSIT]	Lead-Free	yes
03-09-2042	Molex	5935-00-483-0259 [WSIT]	RoHS	yes
03-09-2092	MOLEX INC.	5935-00-137-4732 [WSIT]	RoHS	yes
03-09-2092	Molex	5935-00-137-4732 [WSIT]	RoHS	yes
03-09-2092	Molex	5935-00-137-4732 [WSIT]	Lead-Free	yes
031-0050-0001	Tyco Electronics	5935-00-973-0558 [WSIT]	RoHS	yes
031-0050-0001	Tyco Electronics	5935-00-973-0558 [WSIT]	ELV	yes
031-9134-001	ITT Interconnect Solutions	5999-00-931-5939 [WSIT]	RoHS	yes
031-9134-001	ITT Interconnect Solutions	5999-00-931-5939 [WSIT]	Lead-Free	yes
031-9134-004	ITT Interconnect Solutions	5999-01-180-0152 [WSIT]	RoHS	yes
031-9134-004	ITT Interconnect Solutions	5999-01-180-0152 [WSIT]	Lead-Free	yes
032-0023-0001	Tyco Electronics	5935-00-912-9575 [WSIT]	ELV	no
032-0023-0001	Tyco Electronics	5935-00-912-9575 [WSIT]	RoHS	no
033-0090-0001	Tyco Electronics	5935-00-615-6889 [WSIT]	ELV	no
033-0090-0001	Tyco Electronics	5935-00-615-6889 [WSIT]	RoHS	no
1-102619-0	Tyco Electronics	5935-01-184-5430 [WSIT]	RoHS	no
1-102619-0	Tyco Electronics	5935-01-184-5430 [WSIT]	ELV	no
1-103167-4	Tyco Electronics	5935-01-451-6928 [WSIT]	ELV	no
1-103167-4	Tyco Electronics	5935-01-451-6928 [WSIT]	RoHS	no
1-103167-7	Tyco Electronics	5935-01-116-8384 [WSIT]	RoHS	no
1-103167-7	Tyco Electronics	5935-01-116-8384 [WSIT]	ELV	no

Figure 6 - Compliance information associated with a weapon system

The material information retrieval described above only shows information on materials that appear in one of the four lists of concern. A second retrieval form allows the user to enter a material name or CAS number and returns a list of parts and/or NSNs known to contain that material. Figure 7 shows the input form and results for a search on beryllium. Note that the retrieval header indicates the number of results found, along with the information that beryllium is one of the materials on the DoD emerging contaminants list.

Over all, the response to the SMART web interface was favorable during the beta test period. The tool was made available on the public Pin Point site in September 2012.

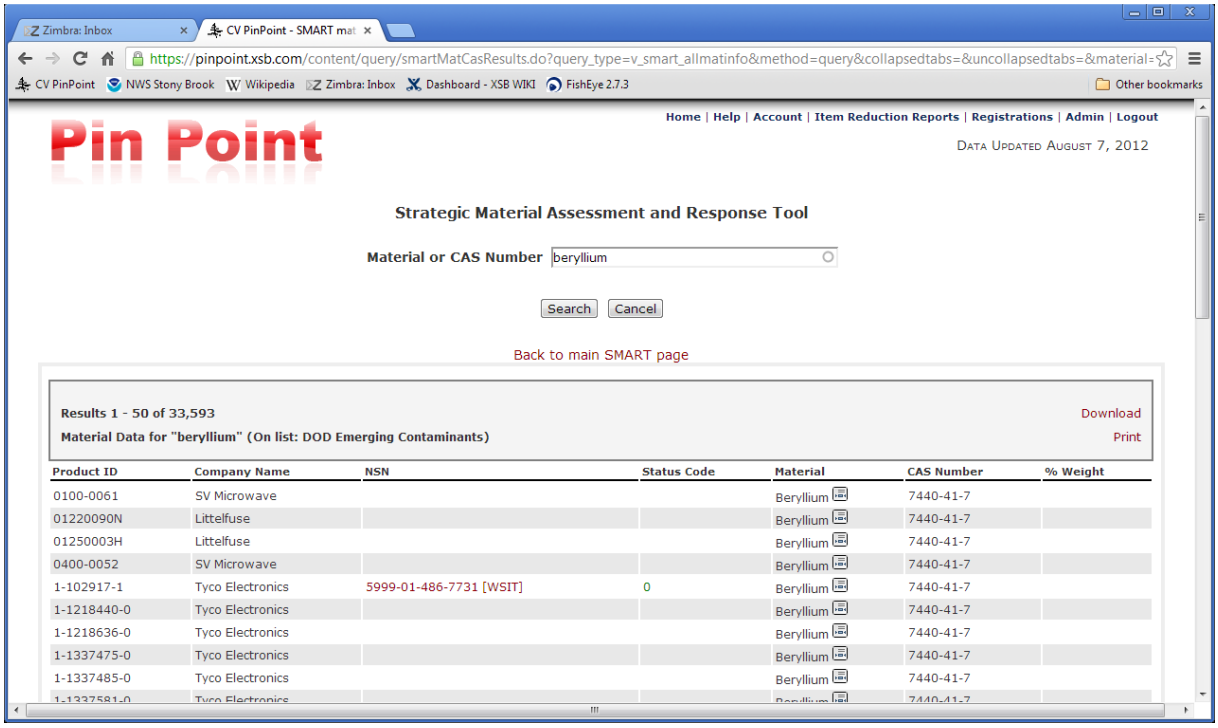


Figure 7 - Retrieval for all parts containing beryllium

Alert System

As mentioned above, we included in the Pin Point interface the ability to register for alerts on regulatory changes. These registrations are carried out using user defined lists of part numbers or NSNs. For example, a user might register for alerts on REACH materials for a given parts list. Then, if materials are added to the REACH list that match any materials associated to parts in the list, the user will get an email alert. In Figure 8, we show an alert subscription pop-up window, which is reached from the the batch maintenance form. Using this form, the user can subscribe or unsubscribe to alerts using the check boxes shown.

Figure 8 - Pop-up window for alert subscriptions

Relationship with HMIRS

There is some overlap in content and functionality between the information supplied by our web interface and the DLA's Hazardous Materials Information Resource System (HMIRS). One of the key differences is source material. HMIRS data is based on a more up-to-date and complete library of MSDS documents. On the other hand, our data contains more information on items such as electronics that do not typically require MSDS, but may contain materials of concern such as lead and beryllium. The interface for HMIRS is geared towards retrieving detailed information for one item at a time, while the Pin Point interface is geared towards obtaining material and compliance information for lists of parts.

There has been some discussion about the possibility of including data from HMIRS into the SMART database, but no decision had been made at the time of this report.

Conclusions

In order to operate in a rapidly evolving regulatory environment, it is necessary to have as much accurate and timely information as possible about items in the supply chain. Often, this information is dispersed in a number of documents of different types. In order to collect material and compliance information from a variety of sources, new technologies for information extraction were combined with existing tools. Two key technologies developed or extended for this project were the extraction of tagged text from PDF documents, and custom annotators for finding material data and other key information in a UIMA environment.

On the whole, the discovery and extraction went well for the available documents. The one exception to this is the extraction of material information from product datasheets. For these documents, there were too many references to materials that were not in the context of constituent materials, making the overall results somewhat unreliable. This situation could be improved with additional effort by adding some level of natural language processing to better evaluate the context of extracted material information.

An important limitation on the completeness of the collected data is the availability of source material. While rich information can be extracted from MSDS documents, only a limited number are publically available via open sources such as the web. More often, manufacturers make them available only to purchasers of the related item.

Regardless of the completeness, the information that was discovered, extracted, and disseminated in the course of this project has proved to be useful to many of the personnel involved in the beta testing of the Pin Point interface. With the functionality now available to all Pin Point users, it is our hope that it will be useful to a wider audience.